MODÉLISATION STOCHASTIQUE - DATA731

Mini-projet 8H sur la mesure d'information à partir de modèles probabilistes

Introduction

La détection de changements est un problème d'intérêt majeur, entre autres applications en :
- cybersécurité (détection d'anomalies ou de cyberattaques, analyse de conformité de signatures numériques),
- économie/finance (rupture de cours boursiers ou de stocks, détection de comportements spéculateurs),
- sciences sociales (recherche de contenus "anormaux" dans les informations associées aux réseaux sociaux, surveillance de foule ou d'attitude comportementale),
- sciences de la terre et atmosphère (surveillance du climat, des volcans, glaciers et forêts),
- protection des villes intelligentes (télésurveillance des infrastructures par caméras),

Les outils associés à un problème de détection de changements sont très génériques en ce sens qu'ils ne dépendent pas en général du domaine d'application d'où proviennent les données : ils dépendent plutôt de la nature de la donnée au sens des interactions entre informations contenues dans la donnée et la présence d'incertitudes.

Du fait de cette généricité, la source des données associée à la série de problèmes définie ci-dessous n'est pas fournie. On se place plutôt le scénario "fictif" suivant : détecter à partir de données hydrologiques annuelles recueillies sur N stations, l'instant associé à la pluie de météorites qui a conduit à la disparition des *dinodroidus* de la planète *Jurassidu* (changement climatique à long terme). Il s'agira de calculer des similarités avant et après le supposé événement, en tenant compte des corrélations éventuelles entre pluviométries inter-stations, afin d'estimer de manière précise, cette date de changement majeur.

Exercice 0 : Introduction et familiarisation au modèle Gaussien

Analyse n°1 :
- Générer N = 100 réalisations d'une Variable Aléatoire (VA) X suivant une loi gaussienne [Normale : $\mathcal{N}(\mu, \sigma^2)$], centrée ($\mu = 0$) et de variance $V = \sigma^2 = 1$.
- Tracer la répartition (histogramme) des valeurs prises par cette VA en divisant la droite réelle en b intervalles. On choisira entre $b \in \{12, 24, 36\}$.
- Normaliser la répartition pour qu'elle soit assimilable à une estimée de densité de probabilité *ddp*.
- Sur cette même figure, représenter la *ddp* théorique f_X et comparer les résultats obtenus.
- Réitérer cette série de manipulations en générant N = 10 000 réalisations de la VA.

Que concluez-vous ?

Analyse n°2 : Reprendre l'Analyse 1 avec une VA suivant une loi gaussienne de moyenne 2 et de variance 9.

Analyse n°3 : Dans cette étude on s'intéresse aux fluctuations des estimateurs empiriques de la moyenne et la variance quand le nombre de réalisations N augmente. Pour cela, nous allons suivre les étapes suivantes :
- Pour chaque valeur de N variant de 100 à 10 000 avec un pas de 50 :
  - Générer N réalisations d'une VA gaussienne avec $\mu = 1$ et $\sigma^2 = 9$.
  - Calculer et stocker les estimateurs Empiriques des moyennes et variances dans deux tableaux nommés Emean et Evar.
- Sur deux figures différentes, tracer les courbes correspondant à Emean et Evar.
- Superposer à ces graphes, des droites horizontales correspondant aux paramètres théoriques μ et σ^2.

Qu'observe-t-on ? Commenter l'évolution des estimateurs empiriques en fonction du nombre de réalisations.

Exercice 1 : Sélection et validation de modèle

Télécharger les données (X_{pluv}). Il s'agit d'une matrice $M = (\mathbf{x}_{i,k})_{1 \leqslant i \leqslant T, 1 \leqslant k \leqslant K} \triangleq (X_1, X_2, \ldots, X_K)$ pour laquelle, la colonne notée ici X_k représente la *variable* dont les échantillons sont des données recueillies à la station n° k ; par exemple les stations X_1 : Annecidu, X_2 : Annemassidu, X_3 : Atlantidu, *etc.*

On considère les 3 couples de variables $Y_1 = (X_1, X_2)$, $Y_2 = (X_2, X_3)$, $Y_3 = (X_3, X_1)$.
- Calculer les matrices de la covariance des vecteurs Y_1 Y_2 et Y_3.
- Que peut-on déduire sur les corrélations existantes entre les pluviométries des 3 stations X_1, X_2, X_3 ?
- Calculer l'histogramme bivarié pour chaque Y_k, $k = 1, 2, 3$. Visualiser ces histogrammes. Normaliser ces histogrammes de manière à ce qu'ils représentent des densités de probabilité (*ddp*).
- Calculer et afficher les *ddp* théoriques Gaussiennes (en utilisant la formule analytique) de mêmes moyennes et covariances que celles, empiriques, calculées ci-dessus. L'hypothèse de distribution Gaussienne des données, est-elle satisfaisante [1] globalement ? L'est-elle localement ? Commentez les résultats.

1. Une étude approfondie de cette question pourra se faire ultérieurement, en utilisant des outils statistiques dédiés, tels que les tests de Kolmogorov-Smirnov.

WARNING : on va maintenant entrer dans le vif du sujet et il est par la suite préconisé de faire d'abord, systématiquement, une analyse mono-variée sur chacune des 3 premières séries d'observations, avant d'implémenter l'analyse multivariée. On écrira les formes simples des entropies relatives et fonctions de vraisemblance associées à une variable aléatoire Gaussienne avant d'implémenter ces fonctions.

Exercice 2 : Analyse globale par entropie relative

On utilise comme mesure de comparaison de variables/vecteurs aléatoires, la divergence de Kullback-Leibler (KL, entropie relative). On rappelle que la divergence symétrique de KL entre deux vecteurs aléatoires $Z_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ et $Z_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ est $\mathcal{D}\left(Z_i, Z_j\right) \triangleq \mathcal{D}\Big(\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \Big)$ est définie par :

$$\mathcal{D}\Big(\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \Big) = \frac{1}{2}\Big(\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)' \left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right) \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) \Big) + \frac{1}{2}\,\mathrm{Trace}\Big(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\Big) - K \quad (1)$$

où $|A|$ et $\mathrm{Trace}(A)$ sont respectivement le déterminant et la trace de la matrice A (le symbole $'$ représente la transposition).

Dans l'hypothèse où le changement s'est produit dans l'intervalle $]t_{m^*}, t_{m^*+1}]$, alors

$$\mathcal{A}(m^*) = \mathcal{D}\Big(\Big(\boldsymbol{\mu}_{[t_1,t_{m^*}]}, \boldsymbol{\Sigma}_{[t_1,t_{m^*}]}\Big), \Big(\boldsymbol{\mu}_{[t_{m^*+1},t_T]}, \boldsymbol{\Sigma}_{[t_{m^*+1},t_T]}\Big) \Big) \quad (2)$$

devrait être maximale du fait des propriétés de divergence de la KL sous hypothèse de validité du modèle Gaussien, où $\Big(\boldsymbol{\mu}_{[t_1,t_{m^*}]}, \boldsymbol{\Sigma}_{[t_1,t_{m^*}]}\Big)$ et $\Big(\boldsymbol{\mu}_{[t_{m^*+1},t_T]}, \boldsymbol{\Sigma}_{[t_{m^*+1},t_T]}\Big)$ correspondent aux paramètres estimés respectivement sur les données avant et après l'instant de changement.

La détection de changement par maximum de divergence KL consiste alors à évaluer des divergences KL sous hypothèse de changement de loi à la date m pour $1 \leqslant m \leqslant T$ et déterminer la valeur m^* correspondant au maximum de divergence.

Calculez les valeurs prises par la divergence $\mathcal{A}(m)$ de l'Eq. (2) lorsque m parcourt l'intervalle $[\![200, T-200]\!]$ (on suppose ainsi que le changement n'a pas eu lieu au début, ni à la fin des observations). Tracez ces valeurs et en déduire m^*, puis les paramètres des lois correspondants :

— avant le changement : $\Big(\widehat{\boldsymbol{\mu}}_{[\![1,m^*]\!]}, \widehat{\boldsymbol{\Sigma}}_{[\![1,m^*]\!]}\Big)$

— après le changement : $\Big(\widehat{\boldsymbol{\mu}}_{[\![m^*+1,T]\!]}, \widehat{\boldsymbol{\Sigma}}_{[\![m^*+1,T]\!]}\Big)$

Exercice 3 : Analyse locale par entropie relative

Mêmes questions que dans l'Exercice 2, mais avec une approche dite par *fenêtre glissante* où

$$\mathcal{A}(m) = \mathcal{D}\Big(\Big(\boldsymbol{\mu}_{[t_{m-r},t_{m-1}]}, \boldsymbol{\Sigma}_{[t_{m-r},t_{m-1}]}\Big), \Big(\boldsymbol{\mu}_{[t_{m+1},t_{m+r}]}, \boldsymbol{\Sigma}_{[t_{m+1},t_{m+r}]}\Big) \Big) \quad (3)$$

où l'on regarde localement autour de la date t_m. Comparer les 2 approches au sens de leurs avantages et inconvénients.

Exercice 4 : Analyse globale par maximum de vraisemblance

Soit $\mathbf{x}_i = (\mathbf{x}_{i,k})_{1\leqslant k\leqslant K}$, les mesures effectuées à la date t_i sur l'ensemble de K stations. La distribution Gaussienne multivariée et sa fonction de vraisemblance correspondante sont notées $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ respectivement.

La fonction de vraisemblance approchée, sous hypothèse d'une distribution Gaussienne multivariée associée aux paramètres empiriques $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ estimés à partir de l'ensemble des mesures effectuées dans un intervalle $[t_m, t_n]$ avec $m < n$ est :

$$\mathcal{L}_{[t_m,t_n]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right) \quad = \quad \left|\widehat{\boldsymbol{\Sigma}}_{[\![m,n]\!]}\right|^{-N/2} (2\pi)^{-NK/2} \exp\left\{ -\frac{1}{2} \sum_{m\leqslant i\leqslant n} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{[\![m,n]\!]}\right)' \widehat{\boldsymbol{\Sigma}}_{[\![m,n]\!]}^{-1} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{[\![m,n]\!]}\right) \right\}$$

où

— $N = n - m + 1$ est le nombre de points de mesure dans l'intervalle $[t_m, t_n]$,

— $\widehat{\boldsymbol{\mu}}_{[\![m,n]\!]}$ est la moyenne empirique estimée dans l'intervalle $[t_m, t_n]$: $\widehat{\boldsymbol{\mu}}_{[\![m,n]\!]} = \frac{1}{N}\sum_{i=m}^{n} \mathbf{x}_i$

— $\widehat{\boldsymbol{\Sigma}}_{[\![m,n]\!]}$ est la covariance empirique estimée dans l'intervalle $[t_m, t_n]$:

$$\widehat{\boldsymbol{\Sigma}}_{[\![m,n]\!]} = \frac{1}{N}\sum_{i=m}^{n} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{[\![m,n]\!]})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{[\![m,n]\!]})'$$

Dans l'hypothèse où le changement s'est produit dans l'intervalle $]t_{m^*}, t_{m^*+1}]$, alors $\mathcal{L}_{[t_1,t_{m^*}]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$ et $\mathcal{L}_{[t_{m^*+1},t_T]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$ sont maximales du fait des estimateurs utilisés : ils correspondent aux estimateurs du maximum de vraisemblance sous l'hypothèse où les données sons issues des mêmes lois Gaussiennes, l'une avant changement et l'autre après changement. La détection de changement par maximum de vraisemblance consiste alors à évaluer la log-vraisemblance approchée sur l'intervalle $[t_1, t_T]$ sous hypothèse de changement de loi à la date m :

$$\mathcal{L}^{[m]}_{[t_1,t_T]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right) = \mathcal{L}_{[t_1,t_m]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right) + \mathcal{L}_{[t_{m+1},t_T]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$$

pour $1 \leqslant m \leqslant T$ et déterminer la valeur m^* correspondant au maximum de $\mathcal{L}_{[t_1,t_T]}$.

Pour $N = 10, 20, 30, \ldots, 100$: Calculer les valeurs prises par la log-vraisemblance approchée $\mathcal{L}^{[m]}_{[t_1,t_T]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$ lorsque m parcourt l'intervalle $[\![200, T-200]\!]$ (on suppose ainsi que le changement n'a pas eu lieu au début, ni à la fin des observations). Tracer ces valeurs et en déduire m^*, puis les paramètres des lois correspondants :

— avant le changement : $\left(\widehat{\boldsymbol{\mu}}_{[\![1,m^*]\!]}, \widehat{\boldsymbol{\Sigma}}_{[\![1,m^*]\!]}\right)$

— après le changement : $\left(\widehat{\boldsymbol{\mu}}_{[\![m^*+1,T]\!]}, \widehat{\boldsymbol{\Sigma}}_{[\![m^*+1,T]\!]}\right)$

Exercice 5 : Analyse locale par maximum de vraisemblance

L'estimation d'une matrice de covariance sur un grand nombre de variables peut s'avérer très délicate d'autant plus que dans beaucoup d'applications, on veut détecter un grand nombre de changements relativement voisins les uns des autres (donc très peu de données de la même nature/loi disponibles). Pour se rendre compte de la difficulté que cette situation occasionne, étudier le comportement de la vraisemblance

$$\mathcal{L}^{[m]}_{p}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right) = \mathcal{L}_{[\![t_{m-p},t_{m-1}]\!]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right) + \mathcal{L}_{[\![t_{m+1},t_{m+p}]\!]}\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$$

basée sur p mesures de part et d'autre de la date cible m et cela, pour différentes valeurs de p. Comparer ces résultats avec l'approche globale.

Exercice 6 : Analyse locale par fusion de maxima de vraisemblances

Une autre alternative est de se restreindre à une série d'analyses locales monovariées (par station) et de trouver un moyen d'agréger les décisions associées aux K estimateurs de maximum de vraisemblance.

Faites une analyse locale monovariée et en déduire les K estimées de la date de changement, leur moyenne, ainsi que la variance autour de cette moyenne (information complémentaire de dispersion).

Exercice 7 : Analyse locale par fusion d'entropies relatives

Les problèmes évoqués dans l'exercice 3 affectent naturellement la KL car ils portent sur l'estimation des matrices de covariance sur un petit nombre d'échantillons. Faire une analyse locale monovariée en suivant la même démarche que l'exercice 4 (utilisation de la divergence KL sur des variables Gaussiennes par station, puis agrégation des estimées).

Problème : projet détection de changements multiples

Télécharger les données (Pixel (3 changements)) et proposer une méthode de détection de changements dans ces données.